IJCCS, Vol.8, No.1, January 2014, pp. 91~100
ISSN: 1978-1520

Analisis Sentimen Twitter untuk Teks Berbahasa Indonesia dengan *Maximum Entropy* dan *Support Vector Machine*

Noviah Dwi Putranti*[1], Edi Winarko[2]
[1] Jurusan Ilmu Komputer, FMIPA UGM, Yogyakarta
[2]Jurusan Ilmu Komputer dan Elektronika, FMIPA UGM, Yogyakarta
e-mail: ***[1]novi_putranti@yahoo.com.au**, [2]ewinarko@ugm.ac.id.

Abstrak

Analisis sentimen dalam penelitian ini merupakan proses klasifikasi dokumen tekstual ke dalam dua kelas, yaitu kelas sentimen positif dan negatif. Data opini diperoleh dari jejaring sosial Twitter berdasarkan query dalam Bahasa Indonesia. Penelitian ini bertujuan untuk menentukan sentimen publik terhadap objek tertentu yang disampaikan di Twitter dalam bahasa Indonesia, sehingga membantu usaha untuk melakukan riset pasar atas opini publik.

Data yang sudah terkumpul dilakukan proses preprocessing dan POS tagger untuk menghasilkan model klasifikasi melalui proses pelatihan. Teknik pengumpulan kata yang memiliki sentimen dilakukan dengan pendekatan berdasarkan kamus, yang dihasilkan dalam penelitian ini berjumlah 18.069 kata. Algoritma Maximum Entropy digunakan untuk POS tagger dan algoritma yang digunakan untuk membangun model klasifikasi atas data pelatihan dalam penelitian ini adalah Support Vector Machine. Fitur yang digunakan adalah unigram dengan fitur pembobotan TFIDF. Implementasi klasifikasi diperoleh akurasi 86,81 % pada pengujian 7 fold cross validation untuk tipe kernel Sigmoid. Pelabelan kelas secara manual dengan POS tagger menghasilkan akurasi 81,67%.

Kata kunci—*analisis sentimen, klasifikasi, maximum entropy POS tagger, support vector machine, twitter.*

Abstract

Sentiment analysis in this research classified textual documents into two classes, positive and negative sentiment. Opinion data obtained a query from social networking site Twitter of Indonesian tweet. This research uses Indonesian tweets. This study aims to determine public sentiment toward a particular object presented in Twitter businesses conduct market.

Collected data then prepocessed to help POS tagged to generate classification models through the training process. Sentiment word collection has done the dictionary based approach, which is generated in this study consists 18.069 words. Maximum Entropy algorithm is used for POS tagger and the algorithms used to build the classification model on the training data is Support Vector Machine. The unigram features used are the features of TFIDF weighting.Classification implementation 86,81 % accuration at examination of 7 validation cross fold for the type of kernel of Sigmoid. Class labeling manually with POS tagger yield accuration 81,67 %.

Keywords—*sentiment analysis, classification, maximum entropy POS tagger, support vector machine, twitter.*

Received December 1st,2013; Revised January 1st, 2013; Accepted January 15th, 2014

1. PENDAHULUAN

Menurut data yang dirilis situs *Semiocast Dot Com* pada 1 juli 2012 jumlah *tweeps* di Indonesia sebanyak 29,5 juta orang [2]. Jumlah tersebut menempati posisi kelima dunia, sedangkan data yang dirilis situs *A World of Tweets Dot Com* menempatkan Indonesia sebagai negara ketiga terbanyak di dunia dalam menulis *tweet* (kicauan), yakni sebesar 11,39% diperoleh berdasarkan rekaman total jumlah *tweet* seluruh dunia sejak November 2010 dari 383 juta profil pengguna Twitter yang dibuat sebelum 1 Januari 2012 [4]. Jumlah pengguna Twitter di Indonesia merupakan pangsa pasar yang menjanjikan. Maka tidak heran berbagai produsen mulai kelas kecil hingga besar berlomba-lomba menggelola potensi ekonomi besar ini agar produk-produk mereka laku di pasaran sekurang-kurangnya menjadi referensi.

Analisis sentimen yang merupakan bagian dari *opinion mining* [3]. Analisis sentimen dilakukan untuk melihat pendapat terhadap sebuah masalah atau dapat juga digunakan untuk identifikasi kecenderungan hal di pasar [7]. Besarnya pengaruh dan manfaat dari analisis sentimen menyebabkan penelitian ataupun aplikasi mengenai analisis sentimen berkembang pesat, bahkan di Amerika kurang lebih 20-30 perusahaan yang memfokuskan pada layanan analisis sentimen [3].

Saat ini Twitter merupakan sebuah indikator yang baik untuk memberikan pengaruh dalam penelitian [9]. Namun masih belum banyak aplikasi dan metode analisa sentimen yang dikembangkan untuk bahasa Indonesia. Faktor-faktor keuntungan tersebut mendorong perlunya dilakukan penelitian analisis sentimen terhadap dokumen berbahasa Indonesia. Penelitian analisis sentimen ini dilakukan untuk mengetahui sentimen publik mengenai sesuatu dengan menggunakan pendekatan dalam *machine learning* yang dikenal dengan nama *Support Vector Machine* dan *Maximum Entropy Part of Speech Tagging* yang dikhususkan pada dokumen teks berbahasa Indonesia dengan fitur *unigram*.

Pemilihan metode klasifikasi *Support Vector Machine* karena memiliki kemampuan generalisasi dalam mengklasifikasikan suatu *pattern*, tidak termasuk data yang dipakai dalam fase pembelajaran metode tersebut [6]. Pendekatan model *Maximum Entropy* (ME) dipilih dalam *Part of Speech* karena terbukti memiliki cara yang sangat efisien untuk mengintegrasikan satu set fitur yang sangat besar dalam model dengan mudah dan telah berhasil digunakan dalam tugas-tugas seperti *Natural Language Processing* (NLP) sebagai bagian dari penandaan pidato [8] atau informasi ekstraksi [5].

2. METODE PENELITIAN

2.1 Pengumpulan data

Pengambilan data dilakukan berdasarkan *query* atas *term* objek pada aplikasi yang terhubung pada Twitter API. Hasil *query* berupa *tweet* kotor baik untuk data pelatihan maupun hasil *query* pengguna mengalami *preprocessing* yang sama.

2.2 Preprocessing

Tujuan dilakukannya *preprocessing* dokumen adalah untuk menghilangkan *noise*, menyeragamkan bentuk kata dan mengurangi volume kata. Tahapan yang dilakukan dari dokumen *preprosessing* dapat dilihat pada Gambar 1.

Gambar 1 Proses *preprocessing* dokumen

Pada tahap *preprocessing* terdiri dari proses *cleaning*, *case folding*, *parsing* dan *filtering.*

1. *Cleaning* adalah proses untuk membersihkan dokumen dari kata-kata yang tidak diperlukan untuk mengurangi *noise* pada proses klasifikasi.
2. *Case folding* adalah proses penyeragaman bentuk huruf serta penghilangan tanda baca. Dalam hal ini hanya menerima huruf latin antara a sampai z.
3. *Parsing* yaitu proses memecah dokumen menjadi sebuah kata dengan melakukan analisa terhadap kumpulan kata dengan memisahkan kata tersebut dan menentukan struktur sintaksis dari tiap kata tersebut.
4. *Filter* bahasa adalah proses untuk memilih *tweet* yang berbahasa Indonesia saja dan jika ditemui kata berbahasa Indonesia tidak baku maka diganti dengan sinonimnya berupa kata baku yang sesuai dengan Kamus Besar Bahasa Indonesia.

2.3 Part of Speech (POS) Tagger

POS tagger adalah sebuah proses untuk memberikan kelas pada sebuah kata. Dalam proses *POS tagger* dilakukan dengan cara *parsing*, kemudian ditentukan kelas tiap kata dengan menggunakan bantuan kamus yang di buat sendiri berdasarkan Kamus Besar Bahasa Indonesia (KBBI) menggunakan metode *Maximum Entropy*. Proses *POS tagging* terbagi ke dalam tiga proses yaitu pemisahan setiap token dalam dokumen dengan pengecekan setiap kata dalam dokumen, mengidentifikasi setiap kata dalam dokumen dengan pemberian jenis kata, pengecekan kata yang belum teridentifikasi terhadap bentuk imbuhan dan akhiran sehingga diperoleh kata dasar.

Berdasarkan aturan linguistik pada kata diperoleh sentimen sementara. Penentuan sentimen dilakukan dengan melihat adanya kata yang mengandung opini baik yang memiliki *polarity* positif maupun negatif dari *tweet* yang sudah dilabeli kelas katanya. Kelas kata yang dipilih adalah kata sifat (*adjective*), kata keterangan (*adverb*), kata benda (*noun*) dan kata kerja (*verb*), sesuai dengan penelitian [3] bahwa keempat jenis kata di atas merupakan jenis kata yang paling banyak mengandung sentimen. Dalam sistem ini jika suatu *tweet* terdapat kata benda (NN) pada sebelum atau setelah kata sifat (JJ) atau kata keterangan (RB) dan kata benda (memiliki *polarity* berlawanan dengan kata sifat atau kata keterangan maka *polarity* yang diperoleh berdasarkan kata sifat atau kata keterangan, karena kata sifat atau kata keterangan memberikan penegasan terhadap kata benda.

2.4 Metode Klasifikasi

Proses pelatihan dilakukan menggunakan *Support Vector Machine* dimulai dengan mengambil data *tweet* bersih dan telah dilabeli kelas sentimennya secara otomatis berdasarkan *lexicon (dictionary)* yang dihasilkan dalam penelitian ini berjumlah 18.069 kata. Kemudian dilakukan proses ekstrak kata unik yang muncul dalam keseluruhan data tersebut. Dari proses ekstrak kata unik tadi diperoleh jumlah *vocabulary.* Selanjutnya adalah menghitung jumlah *tweet* pada keseluruhan data tersebut. Jumlah *tweet* pada masing-masing kelas sentimen dihitung pada tahapan pembobotan. Operasi kernel dilakukan terhadap seluruh data pelatihan sebagai model klasifikasi. Untuk mendapatkan hasil klasifikasi terbaik, diujikan menggunakan fungsi linear dan tiga kernel yang berbeda, yaitu *polynomial*, RBF dan Sigmoid. Dengan memanfaatkan model klasifikasi yang telah dihasilkan maka dilakukan pencocokan *term* tiap *tweet* dengan *term* pada model klasifikasi. Tahap selanjutnya adalah menghitung bobot dari tiap *term* pada setiap kelas sentimen menggunakan *Feature Frequency (FF) Feature Presence (FP)* atau *Term Frequency – Inverse Document Frequency (TFIDF).*

2.5 Pembobotan

Untuk penelitian ini fitur yang digunakan adalah *unigram* dengan pembobotan menggunakan TF, TP dan TF-IDF, kata dan simbol direpresentasi ke dalam bentuk vektor, dimana tiap kata atau simbol dihitung sebagai satu fitur. Adapun perhitungan bobot yang digunakan adalah:

1. *Feature Term Frequency (TF)*

$$\vec{d} := (n_1(d), n_2(d), \ldots n_m(d)) \quad (1)$$

2. *Feature Term Presence (TP)*

$n_i(d)$ = **1** , jika fitur fi ada di dokumen d (2)

$n_i(d)$ = **0** , jika fitur fi tidak ada di dokumen d (3)

3. *Term Frequency – Inverse Document Frequency (TF-IDF)*

$$n_i \text{ (d)} = \text{d}f_i \text{ . Log D/d}f_i \quad (4)$$

Dimana :

df_i adalah banyaknya dokumen yang mengandung fitur i (kata) yang dicari
D adalah jumlah dokumen

Setelah perhitungan bobot tiap *term* dilakukan, selanjutnya proses penentuan kelas sentimen yang memberikan argumen maksimum dengan membandingkan nilai dari ketiga kelas sentimen tersebut.

2.6 Validasi dan Evaluasi

Proses validasi menggunakan 10-*fold cross validation*, dimana data dibagi secara acak menjadi 10 bagian data dengan jumlah yang sama. Sehingga dilakukan proses validasi sebanyak 10 kali secara berulang. Sedangkan untuk tingkat kebenaran proses klasifikasi ditabulasikan dalam suatu Tabel yang disebut *confusion matrix.*

3. HASIL DAN PEMBAHASAN

Data kotor yang digunakan berjumlah 81.885 *tweet* dari penelitian Aliandu [1]. Setelah melalui proses *preprocessing* dan *POS tagger* terdapat 44.006 data bersih yang terdiri dari 12.939 *tweet* positif, 12.654 *tweet* negatif dan 18.413 *tweet* netral yang digunakan sebagai data pelatihan untuk membangun model klasifikasi.

Dalam tahap *preprocessing* menggunakan algoritma *Maximum Entropy* dengan memanfaatkan korpus yang terdapat pada kamus kata yang dibuat dalam penelitian ini berjumlah 18.069 kata. Pengujian pertama dilakukan dengan mengumpulkan *tweet* bersih yang sudah dianotasikan berdasarkan *emoticon* dari hasil penelitian [1]. *Test set* yang dibangun menggunakan 300 *tweet* dari hasil anotasi dengan *emoticon* yang terdiri dari 100 *tweet* positif, 100 *tweet* negatif dan 100 *tweet* netral, kemudian dibandingkan dengan POS *tagger* diperoleh tingkat akurasi yang dapat dilihat pada Tabel 1.

Tabel 1 Perbandingan akurasi tweet beranotasi emoticon(a) dengan POS Tagging (b)

	***Emoticon* (a)**				
	Positif	**Negatif**	**Netral**	**Jumlah**	**%**
Positif	27	14	29	70	38,57 %
Negatif	11	45	14	70	64,28%
Netral	62	41	57	160	35,63 %
Jumlah	100	100	100	300	**43,00 %**
	***POS Tagging* (b)**				
	Positif	**Negatif**	**Netral**	**Jumlah**	**%**
Positif	55	6	9	70	78,57 %
Negatif	10	43	17	70	61,43%
Netral	8	5	147	160	91,88 %
Jumlah	73	54	173	300	**81,67 %**

Hasil anotasi *tweet* yang dilakukan secara manual berdasarkan *emoticon* yang diperlihatkan pada Tabel 1 (a) penjumlahan kelas positif, kelas negatif dan kelas netral yang dikelompokkan dengan benar berbanding jumlah keseluruhan kelas baik positif, negatif dan netral maka diperoleh akurasi sebesar 43,00 %. Sedangkan Tabel 1 (b) memperlihatkan hasil anotasi dengan POS *Tagging* dari keseluruhan kelas *tweet* yang berjumlah 300 *tweet* tersebut menghasilkan 73 *tweet* positif, 54 *tweet* negatif dan 173 *tweet* netral. Berdasarkan penjumlahan kelas positif, kelas negatif dan kelas netral yang dikelompokkan dengan benar berbanding jumlah keseluruhan kelas baik positif, negatif dan netral maka diperoleh akurasi sebesar 81,67 %.

Pengujian akurasi dilakukan dengan jumlah k = 1 sampai 10 menggunakan keempat jenis *kernel*, hasilnya terdiri dari akuransi dan lama durasi pemrosesan yang dapat dilihat pada Tabel 2.

Tabel 2 Perbandingan hasil akurasi klasifikasi tweet dengan Support Vector Machine

Fold	**Tipe *Kernel***	**Linear**	**RBF**	**Polinomial**	**Sigmoid**
1	**waktu proses (detik)**	4403 detik	1723 detik	43593 detik	1695 detik
	akurasi (%)	78,57 %	86.72 %	69,75 %	86,72 %
2	**waktu proses (detik)**	4009 detik	1622 detik	92518 detik	1684 detik
	akurasi (%)	78,59 %	86,61 %	71,02 %	86,61 %
3	**waktu proses (detik)**	4656 detik	1677 detik	75097 detik	1828 detik
	akurasi (%)	78.60 %	86.73 %	71,64 %	86,71 %
4	**waktu proses (detik)**	23826 detik	1630 detik	61302 detik	1715 detik
	akurasi (%)	78,58 %	86,65 %	68,35 %	86,58 %
5	**waktu proses (detik)**	4072 detik	1634 detik	44893 detik	1684 detik
	akurasi (%)	78,57 %	86,57 %	65,01 %	86,62 %
6	**waktu proses (detik)**	4319 detik	1626 detik	52447 detik	1674 detik
	akurasi (%)	78,57 %	86,64 %	60,55 %	86,71 %
7	**waktu proses (detik)**	8886 detik	1632 detik	66273 detik	**1688 detik**
	akurasi (%)	78,58 %	86,39 %	61,99 %	**86,81 %**
8	**waktu proses (detik)**	4437 detik	1631 detik	63019 detik	1694 detik
	akurasi (%)	78,59 %	86,60 %	60,40 %	86,80 %
9	**waktu proses (detik)**	6505 detik	1631 detik	87654 detik	1678 detik
	akurasi (%)	78,54 %	86,61 %	61,52 %	86,62 %
10	**waktu proses (detik)**	4860 detik	1734 detik	114303 detik	1703 detik
	akurasi (%)	78,58 %	86,54 %	61,54 %	86,76 %

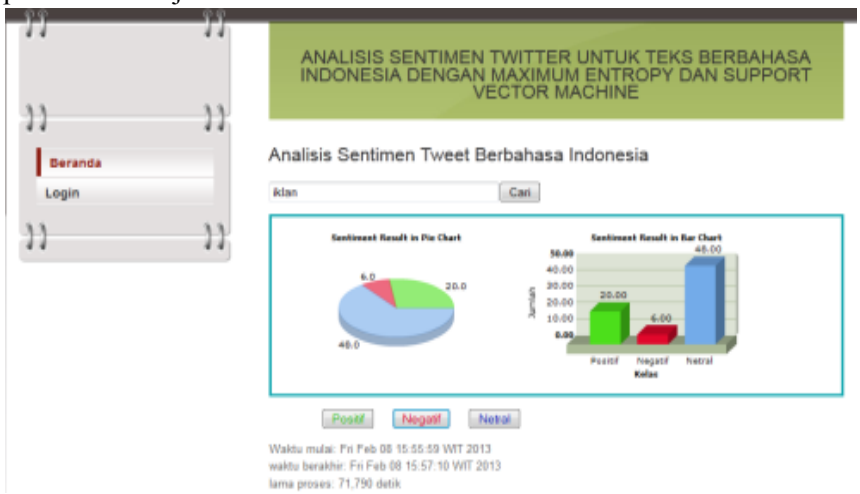
Berdasarkan pengujian pada Tabel 2 diatas dapat diketahui bahwa klasifikasi *tweet* yang memiliki akurasi paling tinggi menggunakan 7 *fold cross validation* pada tipe kernel Sigmoid sebesar 86,81 % dengan waktu proses 1688 detik. Namun, klasifikasi *tweet* terendah pada tipe kernel Polynomial sebesar 60,55 % dengan waktu proses 66273 detik menggunakan 6 *fold cross validation.*

Percobaan aplikasi dilakukan terhadap *query* pada berbagai kategori *term* objek. *Term* objek yang diuji antara lain operator seluler, iklan, Telkomsel, dan Indosat, *term* objek umum yang digunakan adalah operator seluler dan *term* objek khususnya adalah Telkomsel dan Indosat.

1. Menampilkan term objek umum “operator selluler”.

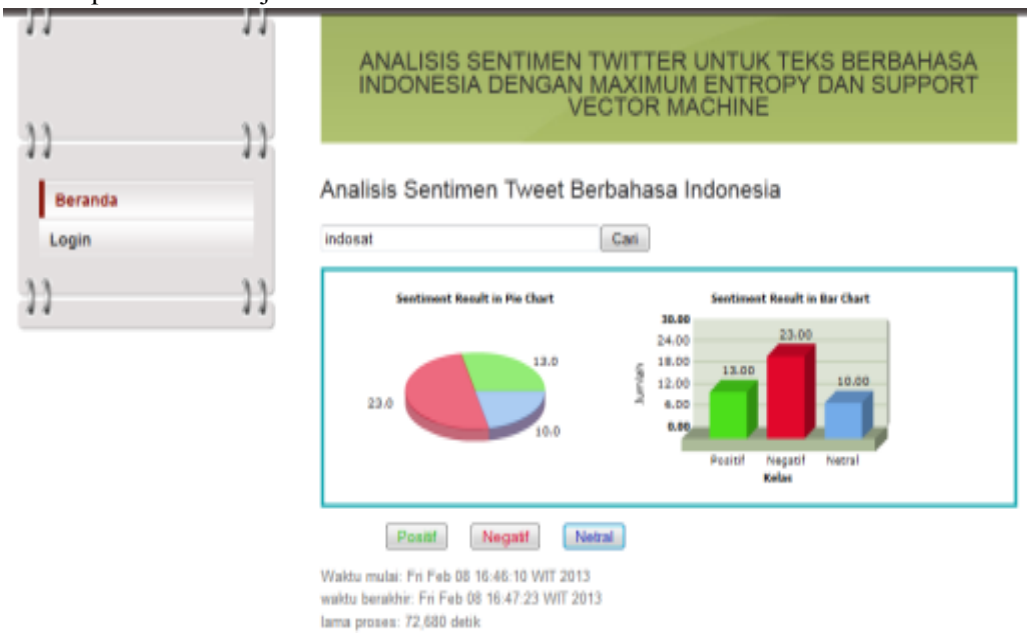
Gambar 2 Tampilan hasil sentimen query “operator selluler”

2. Menampilkan term objek umum “iklan”



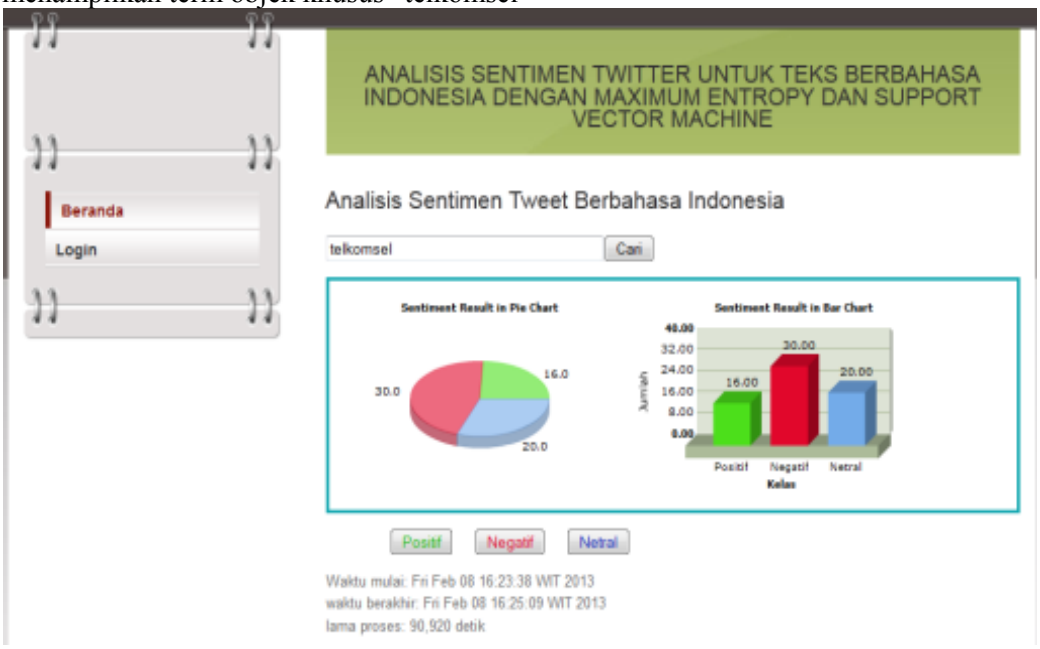
Gambar 3 Tampilan hasil sentimen query kata “iklan”

3. Menampilkan term objek khusus “Indosat”



Gambar 5 Tampilan aplikasi hasil query sentimen kata “Indosat”

4. menampilkan term objek khusus “telkomsel”



Gambar 4 Tampilan aplikasi hasil query sentimen kata “Telkomsel”

Pada Tabel 3 hasil akurasi *query* pada kata "operator selluler" penjumlahan kelas positif yang dikelompokkan dengan benar berjumlah 56, kelas negatif yang dikelompokkan dengan benar berjumlah 5 dan kelas netral yang dikelompokkan dengan benar berjumlah 9 berbanding jumlah keseluruhan kelas baik positif, negatif dan netral berjumlah 73 maka diperoleh akurasi sebesar 95,89%.

Tabel 3 Hasil akurasi percobaan aplikasi untuk query kata "operator selluler"

Kelas	Positif	Negatif	Netral	Jumlah	Akurasi (%)
Positif	56	0	1	57	98,25 %
Negatif	0	5	2	7	71,43 %
Netral	0	0	9	9	100 %
Jumlah	56	5	12	73	**95,89 %**

Pada Tabel 4 hasil akurasi *query* pada kata "iklan" penjumlahan kelas positif yang dikelompokkan dengan benar berjumlah 17, kelas negatif yang dikelompokkan dengan benar berjumlah 5 dan kelas netral yang dikelompokkan dengan benar berjumlah 45 berbanding jumlah keseluruhan kelas baik positif, negatif dan netral berjumlah 74 maka diperoleh akurasi sebesar 90,54%.

Tabel 4 Hasil akurasi percobaan aplikasi untuk query kata "iklan"

Kelas	Positif	Negatif	Netral	Jumlah	Akurasi (%)
Positif	17	1	2	20	85,00 %
Negatif	1	5	0	6	83,33 %
Netral	1	2	45	48	93,75 %
Jumlah	19	8	47	74	**90,54 %**

Pada Tabel 5 penjumlahan kelas positif yang dikelompokkan dengan benar berjumlah 14, kelas negatif yang dikelompokkan dengan benar berjumlah 26 dan kelas netral yang dikelompokkan dengan benar berjumlah 16 berbanding jumlah keseluruhan kelas baik positif, negatif dan netral berjumlah 66 maka diperoleh akurasi sebesar 84,85 %.

Tabel 5 Hasil akurasi percobaan aplikasi untuk query kata "telkomsel"

Kelas	Positif	Negatif	Netral	Jumlah	Akurasi (%)
Positif	14	1	1	16	87,50 %
Negatif	3	26	1	30	86,67 %
Netral	2	2	16	20	80,00 %
Jumlah	18	30	18	66	**84,85 %**

Pada Tabel 6 penjumlahan kelas positif yang dikelompokkan dengan benar berjumlah 12, kelas negatif yang dikelompokkan dengan benar berjumlah 18 dan kelas netral yang dikelompokkan dengan benar berjumlah 9 berbanding jumlah keseluruhan kelas baik positif, negatif dan netral berjumlah 46 maka diperoleh akurasi sebesar 84,78 %.

Tabel 6 Hasil akurasi percobaan aplikasi untuk query kata "indosat"

Kelas	Positif	Negatif	Netral	Jumlah	Akurasi (%)
Positif	12	1	0	13	92,31 %
Negatif	3	18	2	23	78,26 %
Netral	1	0	9	10	90,00 %
Jumlah	16	19	11	46	**84,78 %**

4. KESIMPULAN

1. Data bersih yang dihasilkan melalui proses *preprocessing* dan POS *Tagging* terdapat 44.006 data bersih yang terdiri dari 12.939 *tweet* positif, 12.654 *tweet* negatif dan 18.413 *tweet* netral, lebih banyak dibandingkan data bersih yang dihasilkan berdasarkan *emoticon*, hanya diperoleh 30.813 data bersih yang terdiri dari 10.271 *tweet* positif, 10.271 *tweet* negatif dan 10.271 *tweet* netral.

2. Pada anotasi *tweet* secara manual penggunaan POS *Tagging* menghasilkan akurasi 81,67 % untuk keseluruhan kelas *tweet.* Sedangkan berdasarkan *emoticon* hanya dihasilkan akurasi 43,00 % untuk keseluruhan kelas *tweet.*

3. Metode *Support Vector Machine* dari aplikasi yang dibangun pada *test set* yang dianotasikan dengan *POS Tagging* menghasilkan akurasi sebesar 86,81 % dengan waktu proses 1688 detik menggunakan 7 *fold cross validation* pada tipe *kernel* Sigmoid.

DAFTAR PUSTAKA

[1] Aliandu, 2012, Analisis Sentimen Tweet Berbahasa Indonesia di Twitter, *Tesis*, Fakultas MIPA, Pasca Sarjana Ilmu Komputer, Universitas Gadjah Mada, Yogyakarta.

[2] Campagne, J.C., Dux, J., Guyot, P. dan Julien, D., 2012, Twitter reaches half a billion accounts more than 140 millions in the U.S., *http://semiocast.com/ publications/2012_07_30_Twitter_reaches_half_a_billion_accounts_140m_ in_the_US*, diakses tanggal 16 Oktober 2012.

[3] Liu, B., 2010, *Sentiment Analysis and Subjectivity. Handbook of Natural Language Processing*, Second Edition, (editors: N. Indurkhya and F. J. Damerau). Chapman and Hall/CRC, USA.

[4] Markdalen, A. dan Zapponi,C., 2012, Top 20 Countries Chart,*http://aworldof tweets. frogdesign.com/* diakses 17 September 2012.

[5] McCallum, A., Freitag, D., dan Pereira, F., 2000, Maximum Entropy Markov Models for Information Extraction and Segmentation, *Proc. ICML 2000,* pp. 591–598, Stanford, California.

[6] Nugroho, A.S., Witarto, A.B. dan Handoko, D. 2003, Application of Support Vector Machine in Bioinformatics, *Proceeding of Indonesian Scientific Meeting in Central Japan*, Gifu-Japan, December 20, 2003.

[7] Pang, B., Lee, L., dan Vaithyanathan, S., 2002, *Thumbs up? Sentiment Classification using Machine Learning*, in *Proceedings of the ACL-02 conference on Empirical methods in natural language processing*, Volume 10, pp. 79–86, Morristown, NJ, USA.

[8] Ratnaparkhi, A., 1996, A Maximum Entropy Model for Part-Of-Speech Tagging, *In Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Philadelphia, 17-18 Mei 1996.

[9] Weng, J., Lim, E. dan Jiang, J., 2010, TwitterRank: Finding Topic-sensitive Influential Twitterers, *WSDM'10,* New York City, New York, USA, February 4–6 2010.